3-2021

Avoiding the Basilisk: An Evaluation of Top-Down, Bottom-Up, and Hybrid Ethical Approaches to Artificial Intelligence

Cole Shardelow
University of Nebraska - Lincoln

Avoiding the Basilisk: an Evaluation of Top-Down, Bottom-Up, and Hybrid Ethical Approaches to Artificial Intelligence

An Undergraduate Honors Thesis

Submitted in Partial Fulfillment of

University Honors Program Requirements

University of Nebraska-Lincoln

by

Cole Shardelow, BA

Philosophy

College of Arts and Sciences

March 9, 2021

Faculty Mentors:

Aaron Bronfman, PhD, Philosophy

Albert Casullo, PhD, Philosophy

Abstract

This thesis focuses on three specific approaches to implementing morality into artificial superintelligence (ASI) systems: top-down, bottom-up, and hybrid approaches. Each approach defines both the mechanical and moral functions an AI would attain if implemented. While research on machine ethics is already scarce, even less attention has been directed to which of these three prominent approaches would be most optimal in producing a moral ASI and avoiding a malevolent AI. Thus, this paper argues of the three machine ethics approaches, a hybrid model would best avoid the problems of superintelligent AI because it minimizes the problems of bottom-up and top-down approaches while maximizing their advantages. After detailing the importance of discussing morality in ASI's and outlining some necessary conditions, the problems of machine ethics will be considered, and arguments for and against the three approaches will be responded to.

Keywords: philosophy, machine ethics, artificial intelligence, top-down, bottom-up, hybrid

Introduction

In 2010, a popular internet thought experiment known as Roko's Basilisk was proposed on the philosophy and artificial intelligence forum *Less Wrong.* The theory centers on an artificial superintelligent system well beyond the computing power of human minds that attempts to create the most moral outcomes. In doing so, the theory claims the superintelligence will want to be created as fast as possible to maximize its moral outcomes, which leads it to the conclusion that it ought to blackmail humans through torture to bring about its creation ("Roko's Basilisk."), a decision that seems counterintuitive for an agent attempting to be moral. While the theory has several glaring flaws, namely the difficulty of blackmailing someone in the past, it nevertheless points out a problem both science fiction writers and artificial intelligence theorists have attempted to grapple with: how to construct an AI that is both very intelligent and aligns with humanity's moral norms.

This thesis focuses on three specific approaches to implementing morality into artificial superintelligence (ASI) systems: top-down, bottom-up, and hybrid approaches. Each approach defines both the mechanical and moral functions an AI would attain if implemented. While research on machine ethics is already scarce, even less attention has been directed to which of these three prominent approaches would be most optimal in producing a moral ASI and avoiding a malevolent AI. Thus, this paper argues of the three machine ethics approaches, a hybrid model would best avoid the problems of superintelligent AI because it minimizes the problems of bottom-up and top-down approaches while maximizing their advantages. After detailing the importance of discussing morality in ASI's and outlining some necessary conditions, the problems of machine ethics will be considered, and arguments for and against the three approaches will be responded to.

Significance

First, rapidly developing machines necessitate research into their morality. In the coming decades, artificial intelligence (AI) will continue to become more prevalent and advanced. Already, self-driving vehicles are driving on open streets to offer taxi services (Boudway). Optimistic estimates for developing artificial general intelligence (AGI), or the point at which, "a machine [can] perform any task that a human can," place its creation at 2030 (Joshi). Thus, AGI's have the capacity to perform tasks at least as well as the average human can. This does not imply that, by definition, AGI's are ethical agents, but it does imply they have the capacity to become ethical agents. More conservative estimates from researchers in the field predict AGI will be realized by the end of the 21st century (Müller and Bostrom). Thus, although both science fiction writers and actual AI researchers frequently overestimate the speed at which AI will develop (Müller and Bostrom), it seems reasonable to believe AI will be almost equal if not equal in ability to humans in approximately 100 to 200 years. If this is the case and AI's begin to make the same decisions humans do daily, they will require a capacity to act in correct or moral ways. Examining moral systems for AI early would provide the best chance for AI researchers to make effective determinations on AGI design, especially because it is a newly developing field (Yudkowsky 1).

Second, machine ethics ought to be researched because of the potential impact of AI, AGI, and ASI on human life. Self-driving cars have already impacted human lives through fatalities (Gonzales). While this does highlight AI's effects on us, these incidents are often due to underdeveloped or malfunctioning technology. Once AI such as self-driving cars become more prevalent and advanced, however, these machines will likely encounter moral dilemmas, such as where a self-driving car must decide between saving its passengers or avoiding pedestrians

(Awad et al. 59-60). This impact becomes more significant when considering AGI, which will not only be responding to the morally complex situations humans often deal with, but may do so free of direct human interference unlike today's self-driving cars. Having only mechanical abilities without the understanding of proper norms would make interacting with the world extremely problematic (Allen et al., "Machine Ethics" 13). Finally, if or when an artificial superintelligence (ASI) is created, its impact cannot be understated. An ASI might have the ability and cleverness to bring about worldwide effects, from ending humanity to destroying the earth itself (Chalmers 10). However, Yudkowsky points out that, as with all AI systems, artificial superintelligence has the potential to bring about revolutionary benefits, from treating diseases to creating even more advanced technology (8). Thus, building moral systems in AI is essential to preventing harmful outcomes and ensuring AI's act effectively.

Minimum Necessary Conditions for an ASI safe for humanity

Important to the creation of artificial superintelligence is a discussion of what conditions the ASI must meet to be a safe machine. This discussion serves to both elaborate on the many problems designers may face when creating ASI's and establish the assumptions under which I will compare top-down, hybrid, and bottom-up moral systems. These necessary conditions are far from exhaustive, and authors argue further constraints on ASI's ought to be made with respect to the topics, actions, and communication strategies these machines can employ. (Barrett and Baum 400; Chalmers 31; Yampolskiy, "Safety Engineering" 390).

First, humans must have the ability to understand the ASI. As superintelligence implies, the ASI may become so complex in its thinking that it cannot be understood by humanity (Yampolskiy, "Unexplainability"). If it cannot be understood, it could act in ways contrary to our current interests without our knowing it. At minimum, then, humans ought to be able to generally

understand both the actions the AI is taking and the rationale behind the actions. The rationale ought to provide a specific enough explanation so that humans can monitor whether an AI is safe for humanity. For example, if an AI performed an action and stated its rationale for the action was that it was dictated by utilitarianism, that would be too simple an explanation to ensure a safe AI. A good rationale would provide calculations of the utilitarian analysis that humans could understand. Some details may be too complex to comprehend, but they should not be necessary to understanding the general plan of the AI.

Second, Humans must have the ability to communicate with the AI. If an AI reaches superintelligence, it should have the capacity to communicate with humanity in some way. This communication could act as a mechanism to help understand the AI. This communication could come about in the traditional human sense through an interactive dialogue using text or vocals or could be as simple as the AI producing a report diagnosing its actions and status.

Third, humans must be able to alter the AI. The AI must have an ability to be altered by humans to prevent actions against humanity. Communicating and understanding the AI acts as an early warning system for humanity to know when to change the AI's plan of action or alter the internal systems of the AI. The ability to alter the AI is necessary for a moral superintelligence because we may deem the AI's moral system so different or out of touch from ours that it is not in our interest to let the AI act in said way.

Arguments Against Moral Machines

While there are many disagreements as to the type of moral framework AI ought to adopt, there are also authors who believe morality should not be implanted in AI altogether. Roman Yampolskiy presents several alternatives to moral machines under a category he calls,

"AI Safety Engineering" ("Safety Engineering" 390). AI Safety Engineering currently centers on confining AI through mechanical or cognitive constraints. Mechanically, AI designers could restrict AI into simulated virtual programs for researchers to collect data and learn more about the real world from the AI. In this scenario, the AI would be able simulate an action as if it were outside the confines of a digital space without the danger of real consequences (Yampolskiy, "Safety Engineering" 390). On the cognitive side, designers might restrict the types of actions an AI can perform. Yampolskiy highlights concepts such as Oracle AI in which the machine can only answer questions from people, thus minimizing the risk the intelligent AI poses ("Safety Engineering" 390).

Yampolskiy makes two main claims as to why AI Safety Engineering should be preferred over moral machines. First, he notes machines will in some way have human-like moral characteristics, and because humans perform immoral actions regularly, machines based or influenced by humans will have similar moral flaws making moral machines unacceptable. Second, Yampolskiy claims that AI being peaceful and abiding to mutually agreed upon laws is sufficient to avoid dangerous AI ("Safety Engineering" 390).

I will provide some responses to the first claim above. Regarding the human-like flaws of AI, while it seems true that all AI will be influenced by humans, this influence can be mitigated to avoid problematic flaws. First, certain types of moral frameworks, such as bottom-up frameworks, mitigate human flaws because the framework is based on learning morality through experience, rather than directly inputting moral codes into the machine (Boyles 189). Thus, an AI could develop their own morality which could differ enough from human morality that it could avoid some of its flaws. Second, because an AI is a separate being from ourselves and uses very different processes to think and interact with the world, it may conceive morality different

from us and allow it to skirt human moral flaws. Thus, the influence of human moral flaws can be mitigated by AGI's and ASI's.

I will also respond to the claim that peaceful and law-abiding machines are sufficiently safe machines. First, the arguments Yampolskiy presents against moral machines are the same ones that could be presented against his peaceful, law abiding AI concept. These rules for peaceful, law abiding conduct will in some way be influenced by humans or AI will perform the rules in a human-like manner. If this is true, AI safety machines will still encounter human errors and will be no better than an AI that acts under moral rules. While it seems true that a law-abiding machine may be easier to design, this ease could come at the cost of not being able to handle moral dilemmas in which a rule is not available or in which two contradict. Conversely, a moral code could allow an aspect of flexibility to cases. Thus, because both morality and strict laws offer flaws, it may be beneficial to combine laws with a moral code to provide both ease and flexibility. This leads to the second argument that moral machines and AI safety engineering machines are not distinct. In other words, it seems that AI safety engineering is still based in morality but referred to differently. Insofar as AI safety engineering is based on inputting norms of behavior such as being peaceful and abiding by laws into machines, these machines will have systems that could easily fall under moral frameworks such as deontology. Thus, not only does Yampolskiy fail to avoid his own arguments against morality, but he also creates a theory that can be considered a moral framework itself.

Next, there are reasons why AI Safety Engineering is ineffective on its own. While this system may be effective for AGI, David Chalmers points out constraining ASI's will be nearly impossible (38). A superintelligent AI, once it realizes it is constrained in a virtual world or other container, will likely have the capacity to escape as long as it has some connection with the

outside world. For example, it could convince humans to release it or use its limited interactions with the world to escape (Chalmers 38). As a result of these problems, one might try to further restrict information or interaction with the outside world from the ASI, but at that point it becomes a machine that cannot be learned from or useful to society. ASI's will thus need some moral framework along with possible constraints to both provide safety to humanity and effectively improve society.

I will also respond to arguments on the infeasibility of machine ethics. Miles Brundage finds many of our moral beliefs are ambiguous and contain exceptions, making it difficult to map proper morality onto AI (356). Further, he points out our morality may be this way because of evolution, where humans use a dual-track model of morality to avoid antisocial behavior. The implications from these statements are that much of our morality is influenced not in logic, but in the makeup of our species and our social structures. Thus, it may be infeasible to map morality onto machines because they both lack the evolutionary design to accommodate moral concepts and because the exceptions within morality are too numerous and complex for machines to adopt (Brundage 356-357). This also implies morality is a social construct and in no way objective, which would make it difficult to construct morality around a machine as well as make it difficult for a machine to understand what morality is in the first place.

Three arguments can be made to address the infeasibility of morality. First, it seems machines may be able to adopt the evolutionary morality of humans if it is most effective or if no alternatives are available. Brundage highlights how recent psychological research shows our morality functions through a dual-track system, where our intuitive, quick moral thinking uses deontology while our deliberate moral thinking uses utilitarianism (357). AI could adopt this moral framework because utilitarianism and deontology could theoretically be inputted into

machines, especially if that machine is superintelligent. Second, assuming morality is socially constructed and difficult to systematize, bottom-up moral systems could skirt these problems. Bottom-up morality functions similarly to how humans develop morality through experiencing moral dilemmas and learning from those around them. Because AGI's and ASI's, by definition, have human level intelligence and can perform any task a human could, they could learn morality's socially constructed rules to understand the ambiguous details and better interact with humans. Brundage may disagree with the definition that AGI allows the machine to learn morality, but it seems clear that intelligence has a key role to play in morality and that it is not merely emotional. If this is true, the ability of AGI's to have an intelligence equal to ours would give it some ability to imitate our morality. However, we may not want machines that act through human-like morality, which is flawed. ASI's using a bottom-up, hybrid, or top-down system could avoid these problems because they have the intelligence and processing power to calculate utilitarian problems or even reason through the ambiguities humans face in moral dilemmas.

Finally, Yampolskiy argues artificial intelligence research and development itself is unethical. First, he notes the potential consequences of AI are too severe to justify development, as AI's ability to self-improve could lead to the overpowering and extinction of humanity (Yampolskiy, "Safety Engineering" 392). Second, developing an AGI (or ASI) would make humans responsible for creating an entity that can experience suffering, which is wrong. Third, because AGI's (and ASI's) would be intellectually developed to at least the level of a human, it would be unethical to subject them to experiments that are inhumane (Yampolskiy, "Safety Engineering" 392).

I will make several responses to Yampolskiy's claim that AI research is too dangerous. First, while mechanical constraints were previously discussed as being imperfect, they still could have some ability to mitigate the power of artificial intelligence over time, making Yampolskiy's worry of human extinction less likely. Second, Yudkowsky points out that just because an artificial intelligence has the *ability* to destroy humanity does not mean it has the *intention* to do so (12). The AI may have plenty of reasons not to destroy humanity or may just be neutral to the idea. Third, this leads to the importance of morality in AI systems. If AI's are designed around an understanding of right and wrong, they will be less likely to have the intention to harm or destroy humanity. Thus, machine ethics allows humanity to reap the benefits of AI while minimizing the probability of danger. Ultimately, the risk of AGI's or ASI's cannot be completely minimized and may still pose large dangers for humanity. The claim that AGI research is too dangerous is a much stronger claim that must also consider the benefits brought about by AG/SI's in addition to the factors that minimize risk above. For example, Yudkowsky notes powerful AI could be a solution to existential risks (8) such as climate change policy or human migration from Earth. While a powerful AI is dangerous, it may not be *too* dangerous to at the very least research potential designs and advancements.

I will also respond to Yampolskiy's arguments on suffering and experimentation. First, Yampolskiy is viewing suffering from a human point of view, but it is unclear whether robots will experience suffering in the same way we do. Suffering or pain for machines could just be an algorithmic result they are instructed to avoid, not an emotional or physical response. Even machines that have the cognitive abilities of humans will be different enough in how they interact with the world that pain may not be an issue. Thus, the problem of suffering could be avoided altogether based on how the machine is designed.

Second, if a machine can experience suffering, it can also likely experience pleasure. While the consideration of creating beings who can suffer is important, this consideration must also be weighed against the benefits these beings could receive from pleasure. Thus, the potential pleasure an AGI or ASI could receive helps to mitigate the potential problem of suffering.

Finally, Yampolskiy's claim that AGI creation is unethical because invasive experiments could be performed on them assumes these experiments will happen at all. It is likely that once individuals realize the AGI has cognitive abilities equal to humans, they will be involved in less invasive experiments. Thus, less invasive experiments through surveys and interviews eliminate the problem of harm through experiments. One could respond by noting this will severely limit the ability of humanity understand AGI's or ASI's. While it may be more difficult to understand them because said experiments are less invasive, humans would still exert some control over AGI's because they initially designed them and, by one of the previous necessary assumptions, can alter them. Thus, if humans can develop an effective design that would steer an AGI, developers could avoid rogue AGI's while sticking to less intense experiments.

In the proceeding pages I will discuss the three prominent types of machine ethics, some specific theories, and arguments for and against them.

Top-Down Machine Ethics

Top-down morality for ASI's centers on one or more moral theories being the framework for what moral rules the machine will follow. The machine will not follow any other moral codes except the ones explicitly prescribed to them. Wallach and Allen find, "the top-down approach to artificial morality is about having a set of rules that can be turned into an algorithm. Top-down ethical systems might come from a variety of sources…" (84). The idea of a top-down approach

also has meaning in engineering terms, where it's defined as reducing tasks to smaller components (Wallach et al. 568). Thus, top-down morality uses broad theories, such as utilitarianism or deontology, as a foundation for specific moral rules machines must adhere to.

As an example of an applied theory, one conception of top-down ethics comes through utilitarianism. The benefit of utilitarianism within a top-down ASI is its ability to derive morality from calculations about costs and benefits (Wallach and Allen 86). However, its main drawback is the sheer amount of calculations that must be done to account for the outcomes of actions. While ASI's would be able to perform calculations much more effectively than humans, it still may not be able to perform all possible calculations. Thus, researchers have outlined steps for implementing utilitarianism in machines. Wallach and Allen describe the four steps as requiring machines to have, "A way of describing the situation in the world, a way of generating possible actions, a means of predicting the situation that would result from an action, [and] a method of evaluating a situation in terms of its goodness or desirability" (87). Each of these steps requires a large amount of processing power that a superintelligence could handle.

Next, I will describe a specific top-down theory called the dual-process model. The dual-process model is a model theorized by moral psychologists and seems to be one of the primary ways humans make moral decisions. According to the dual-process model, humans can use either a quick firing intuitive ethical calculation or a slower, more deliberative process (Cushman et al. 5). This psychological theory has found deontological reasoning comes from the intuitive process while utilitarian reasoning comes from the deliberative process (Greene 699). This is not to say that one track is inherently better because it is deliberative or not. This theory also does not imply one does or must think through a deliberative, utilitarian process when they have time to deliberate. It does imply, however, that when there is little time to deliberate, the intuitive

track is often used. Thus, the name dual-process stems from the deliberative and intuitive cognitive processes within humans that use utilitarian or deontological ethics respectively (Brundage 357). When having the time and ability to thoroughly analyze a moral dilemma, our brains are more able to use the deliberative process and make judgements based on the utility of the decisions (Cushman et al. 5). The sheer amount of internal calculating that utilitarianism forces a human to do necessitates a deliberative process. Conversely, humans often do not have sufficient time to make important moral decisions, so the intuitive process ensures a decision can be made. This process uses deontological rules which require much less calculating than a typical utilitarian problem. In fact, Cushman et al. reveals, "imposing a cognitive load slowed down characteristically consequentialist judgements, but had no effect on characteristically deontological judgements" (5). Thus, deontology acts as an effective heuristic for moral judgements when one is cognitively strained. The dual process model therefore allows for flexible moral decision making which could be an asset to AI.

It is also important to consider why an artificial superintelligence would benefit from a dual-process model. One might object that because the ASI is super intelligent it would have no need for a model which is meant to compensate for cognitive load. However, it may still be necessary for an ASI to account for cognitive load based on how it is used. For example, an ASI Oracle machine (Yampolskiy, "Safety Engineering" 390) could be designed to field hundreds of questions at a time, using much of its processing power to do so. A less straining intuitive process would thus be useful for allowing an ASI to continue multitasking while also providing moral decisions. Next, some decisions may be too short for even superintelligent machines to respond to. Suppose an ASI had to respond to a terrorist's bomb threat asking a government for an exorbitant sum of money, contraband, and weapons in exchange for not killing a certain

number of civilians. The terrorists know the government is using an ASI to make the decision, so they give the government a minute or less to respond to put it under pressure. While it is unknowable at this point how powerful a superintelligent machine will be, it seems reasonable that even an ASI would have trouble mapping out the utilitarian implications of these types of dilemmas in just a minute. Thus, an intuitive deontological system would be invaluable in the algorithm of an ASI, as it would likely respond quicker and more effectively than humans.

There are several responses that I will make to the dual-track model. First, whether it be deontology or another rule-based system such as Isaac Asimov's three robot laws, each system will likely encounter laws that conflict with each other and are thus unable to be followed during decision making. Wallach and Allen realize, "In the context of the real world, seemingly straightforward rules can also turnout to be impossible to follow…conflict between them can cause a deadlock" (93). One could respond to defend a rule-based system because of the processing power of an ASI. An ASI may be able to create a priority list of hundreds or thousands of rules, allowing it to avoid conflict between them. The only limitation of this solution is increasing the complexity of the system and requiring more time to process an answer.

Next, the discrepancy between a utilitarian and deontological system within one machine produces problems with moral consistency. While much of the ASI's decisions within a dual track system will likely use the deliberative utilitarian process, it will sometimes have to switch to making decisions through a deontological lens. If humanity is using ASI's to work on important political, scientific, etc. endeavors, which we likely will, it will be difficult for stakeholders and society at large to expect consistent types of decisions from the ASI. Wallach and Allen point out, "Even given a set of comprehensive, nonconflicting rules, consecutive

repetition of one or more rules can lead to undesirable results… without consideration of the whole process over a period of time" (93). The problem Wallach and Allen highlight is especially true if two conflicting moral views are used at varying points of an ASI's life. A lack of consistency for an ASI could be catastrophic to humanity, as moral consistency is important for humans to structure their expectations and plans. Further, because ASI's will likely deal with very significant societal matters, a significant number of people's expectations and plans will be subverted compared to a normal human.

A potential defense against moral inconsistency could be to point out humans function this way all the time. The dual-track process is itself based in human psychology and an average person would likely not have their expectations or plans significantly undermined by another's use of a different moral view.

I will respond to this by noting our expectations for consistency ought to be set much higher for an ASI, so we should not see humanity's moral inconsistencies as sufficient to justify an ASI's. It may be true that people either do not notice or do not mind that their expectations or plans are shifted because of another's moral inconsistencies, but they may still occur. An ASI may also have to make significant scientific or other critical decisions which require precision and consistency beyond that of a human, so it must offer moral consistency to complete those tasks.

An additional rebuke of moral inconsistency could be that the AI would already factor the inconsistency into their calculations. The AI would then make decisions based both on what track is used and whether that track would avoid disrupting individuals' expectations or plans in the long term.

I will respond to this by noting the AI factoring in inconsistencies could harm its ability to act morally, which is a general problem for a dual track superintelligence. The additional factor of consistency may force the AI to make a decision that may be worse than a separate decision that would be better but would create inconsistency.

Bottom-Up Machine Ethics

Bottom-up approaches to machine ethics follow human-like patterns to learning morality while offering more freedom than top-down approaches. Boyles furthers, "Bottom-up options…employ evolutionary, learning, or developmental methodologies. Such approach enables machines to learn ethically-related concepts, for instance, via interacting with other agents in their respective environments" (189). Thus, bottom-up approaches center around learning morality through experience while simultaneously lacking any built-in directions, as well as lacking any guarantee AI will accept external directions communicated to them, unlike top-down approaches.

Several current machine processes have been noted as possible routes to implementing bottom-up machine ethics. One route is through Alife and evolutionary algorithms, which are robotic analogues of natural evolution. Alife involves producing an evolutionary struggle for machines within a computer simulation. A current example of this technique involves game theory, where simple AI develop the best way to interact with other AI in a simulation by cooperation or defection. The problem with Alife is its lack of realistic interactions within the virtual world, making it difficult to translate learned robot skills into the real world (Wallach and Allen 100). More tangible benefits for bottom-up artificial intelligences could come through evolutionary algorithms based in physical space. Although this bottom-up strategy is not ethical in nature, one current example discussed by Wallach and Allen is where roboticists have robots

complete tasks and give them a score based on their effectiveness. The highest scoring robots then combine their machine parts with other high scoring robots to simulate reproduction and mutation and the process continues (Wallach and Allen 100-101). An AI in the future may act similarly when trying to determine the most ethical outcomes. Although it will likely make moral mistakes initially, each chance to deliberate and act on a moral situation allows the AI to better learn and evolve from previous cases.

Other bottom-up strategies focus on learning similar to child development as opposed to evolutionary strategies. One such strategy is quest ethics, which tasks machines to achieve rational goals and, using neural networks similar to our brains, make connections between the actions they take to reach those goals and broader ethical strategies (Wallach and Allen 109-110). Quest ethics has the added benefit of guiding machines toward goals already deemed rational to humans, thus providing a method to lightly steer bottom-up machines with human values. The piece that distinguishes quest ethics from a top down strategy is the ability of the AI to learn its own criteria for action and goals as opposed to them being imposed by humans. Thus, even though moral rules may be found within bottom up machines, they are discovered instead of implanted. Regardless of evolutionary or developmental strategies, bottom-up approaches center around repetitive actions to generate improved moral decision-making skills.

There are several benefits to implementing artificial superintelligences with bottom-up approaches. First, bottom-up approaches offer a better ability for ASI to develop new moral reasoning previously not thought of by humans. After all, because of ASI's completely different approach for thinking and interacting with the world, ASI's will likely develop a unique set of moral behaviors, which need not be a difficulty but a benefit (Wallach et al. 243). This ability differs greatly from top-down approaches which are necessarily restricted in how they can

deliberate on and enact moral decisions. Thus, the two positive effects gained from this benefit come in creating an ASI which potentially acts more moral than humans while also providing humans new ethical theories to study and learn from.

The second main benefit of bottom-up ethics is that it makes good use of the processing power of an artificial superintelligence. While top-down ASI's will develop a more comprehensive understanding of the moral theory they are assigned with than humans, they have inherent limits to their moral understanding because of their restricted framework. Conversely, bottom-up designs, within the current technology of AI, are limited by their ability to draw connections between decisions they make and a broader moral scheme (Allen et al., "Prolegomena" 258). The creation of an ASI would traverse that problem and allow it to develop a sophisticated moral system. Thus, a bottom-up system would provide more utility in its processing power by developing new moral schemes with guidance from humans in some way.

Several glaring problems face bottom-up approaches to ASI, even with their ability to process information at superhuman levels. First, bottom up ASI lacks layered architecture designed to protect humans from a malicious ASI. Wallach and Allen explain layered architecture as, "core restraints [that] might be built into foundational layers of the computer platform that are inaccessible to those parts of the computer that learn and revise the structures that process new information" (111). Assuming it would even be possible to create an untouchable foundation within an ASI, while top-down approaches would contain layered architecture due to their moral framework, bottom-up approaches lack it simply by virtue of their defined design. Because bottom-up frameworks are based around developmental moral learning in as free a manner as possible, the only layered architecture they might possess would center around *how* the ASI learns and develops. The central problem this leads to is an ASI that could

develop moral codes both neutral and hostile towards humanity (Brundage 361-362). This, combined with the ASI's intelligence, could cause extreme damage to humanity. Thus, bottom-up approaches could fail their central purpose of making AI compatible with humanity by instead making them deadly.

Next, a bottom-up approach could create an ASI that differs widely from us. Because the ASI would be the least restricted in its pursuit of a unique morality compared to the top-down and hybrid approaches, it may create a type of morality completely unrecognizable to us (Brundage 361). This could occur even if the ASI learned morality from ethically good humans, and even if the morality could technically be understood. It may simply be a moral form which humanity would be unwilling or unable to use themselves. The effect of this problem, other than a morality potentially hostile to humans, is a lack of trust that builds up between an ASI and humanity (Brundage 362). Conflict between an ASI and humanity would be potentially catastrophic, and distrust would only build to such a relationship. Thus, while a bottom-up approach could be more tailored to human morality through proper experiential learning, it lacks the safety rails that other ethical approaches offer and could allow for a rogue ASI.

A further problem for bottom-up technology is that it may not be able to achieve a moral system due to how it is designed to learn. Current bottom-up strategies, such as Alife and quest ethics, center on repetitive learning focused on achieving one goal, but it is uncertain whether these mechanical processes can translate into moral learning. First, even if an ASI could focus on achieving multiple goals because of its processing power, it may not be able to achieve a proper moral system. Wallach and Allen find, "Human morality is dynamic…[AI] will need the capacity to dynamically negotiate or feel their way through to elevated levels of trust with other humans or computer systems with which they interact" (113). ASI's may not be able to feel their

way through trust with humans and vice versa because of how we function. While humans have various emotional states that affect how we trust others, ASI's through a bottom-up approach would try to approach trust through mechanical means that might be incompatible with our thinking. This is also complicated through the bottom-up approach because of its goal-oriented process. The goal of achieving a relationship with a human may be done through immoral or amoral means.

Finally, an ASI using bottom-up morality could easily be trained to be immoral. While it is true both top-down and hybrid methods could produce immoral ASI's by putting immoral frameworks in them, they at least have a framework that allows them to be controlled. Conversely, even if a benevolent engineer wanted to create a morally good bottom-up ASI, it would be difficult to ensure that would happen (Brundage 361). First, because ASI's would require real experiences to develop their moral understanding, they could interact or learn with individuals who teach them immoral behaviors. Second, much like any human might do, a bottom-up ASI may not listen to any one individual who attempts to train it, making moral training difficult to administer.

Hybrid Machine Ethics

Hybrid machine ethics takes its name from its mixture of top-down and bottom-up processes. While top-down approaches limit moral choice through a restrictive framework, bottom-up approaches conversely provide large amounts of moral freedom by requiring the machine to learn morality through experience (Allen et al., "Artificial Morality" 153). Hybrid machine ethics provides a convergence of these, typically by having a lighter moral framework than top-down processes while still allowing the experiential learning of bottom-up processes.

One well known moral theory often cited as a potential hybrid process is virtue ethics. Beavers finds:

> The top-down approach, is directed externally toward others…as a necessary restraint on one's desire with the effect of… promoting liberty and the public good…Bottom-up developmental approaches…can precipitate where, when, and how to take action, and perhaps set restraints on the scope of theory-based approaches…virtue ethics would seem after all a good candidate for implementation. (336-337)

Next, virtue ethics may have a robotic analogue in connectionism that could be implemented in machines. Connectionism, also known as artificial neural networks, is a form of AI in which the machine performs basic tasks and is able to make connections between the tasks to generate categories or patterns (Wallach and Allen 121). For instance, neural networks have been used to connect text to spoken phrases so machines can read text aloud (Wallach and Allen 122).

I will now further describe the hybrid approach and its relationship to overarching principles. At its highest level, top-down principles cannot be modified by bottom up reasoning in a hybrid system. If a top down structure was implanted and focused on rules, utility, or a particular virtue, those could not be modified. However, this does not mean every hybrid AI produced would act the same if given the same top-down principle. The principle, especially for virtues, provides a vague enough skeleton for the AI to learn through experience the best way to understand and achieve the principle. this does not mean that the topmost principles cannot be redirected in fundamental ways. Wallach and Allen find hybrid machines, "[include] top-down values informed by cultivated implicit values and a rich appreciation of context" (118). A hybrid approach builds its own lower tiered values to understand and attain its ultimate principle.

Additionally, this does not imply hybrid approaches are simply a form of causal reasoning to the topmost principle. The cornerstone of the hybrid approach is its development of practical wisdom to create larger values through experience and understanding context, which is why this approach is often illustrated through virtue ethics (Wallach and Allen 118). Finally, although practical wisdom is a type of moral reasoning typically seen in virtue ethics, it can be applied to other top-down principles as well. Hursthouse and Pettigrove point out, "Even many deontologists now stress the point that their action-guiding rules cannot, reliably, be applied without practical wisdom…the capacity to recognize, in any particular situation, those features of it that are morally salient." This practical wisdom could also be applied to utilitarian principles. For instance, a hybrid machine could, through practical wisdom, develop its own rules within a utilitarian structure to build its own rule utilitarian morality.

In general, the hybrid approach to machine ethics offers the best type of morality of the three approaches because it best mitigates the problems faced in the other approaches while optimizing their benefits. I will now describe the hurdles the hybrid approach overcomes as well as rebuke arguments against it.

There are several benefits to connectionist virtue ethics approaches. One main benefit is its ability to be both predictable and flexible by implementing top-down and bottom-up approaches respectively. In other words, the hybrid approach can combine the predictable principles of the top-down approach and the flexible learning of the bottom-up approach to better effect, on balance, than the two approaches alone. This would allow a machine to work within the confines of human behavior while using its flexible bottom-up abilities to learn new, potentially better moral behaviors (Wallach and Allen 122). This will likely mean a hybrid approach will be worse than a top-down approach in terms of flexibility, as it is less rigid.

Conversely, it is not as flexible as a bottom-up approach because of its defined topmost principle. However, if the goal of a moral AGI or ASI is the avoidance of harm to humanity, the hybrid approach secures this most effectively because it is not too rigid or too flexible.

Next, connectionist virtue ethics could overcome competing principles, unlike top-down approaches. As previously noted, top-down approaches suffer from conflicting principles that arise within frameworks such as deontology or rule utilitarianism. A hybrid model would likely encounter this too, but its solution is much more easily implemented than through a top-down approach. A top-down approach would require manual implementation of a rigid tier system to prioritize principles, and this solution still would not be enough to account for the infinite moral dilemmas an ASI would face. A top-down approach may simply lead to an ASI not acting without a proper principle that could apply in a case. Alternatively, a top-down approach may use a vaguer rule that can apply to many cases, but may lack the practical wisdom to apply it effectively given the context. On the other hand, a hybrid system is much more flexible and adaptable to cases, allowing a hybrid ASI to not only respond to a new dilemma more effectively because of its use of practical wisdom, but also learn from the dilemma and better understand which principles to prioritize. This learning would come about through a better understanding of what Hursthouse and Pettigrove call the, "morally salient" features of the situation, potentially leading to the adoption of new principles or values that apply to the situation. Thus, hybrid approaches may be able to both prioritize more effectively and, unlike top-down approaches, learn what is most moral to prioritize.

Additionally, connectionist virtue ethics better prevents an ASI from learning immoral behaviors by having safety rails. These safety guards come in how the top-down and bottom-up approaches interact. A benevolent person who is morally imperfect may inadvertently implant a

hybrid ASI with immoral principles towards certain groups of people, but its bottom-up system may be able to at least minimize those principles by socializing with everyday decent people. Conversely, a hybrid ASI that learns bad behaviors through socialization and experience could have these behaviors subverted by proper moral principles implanted on them. Of course, an ASI could be created that has both immoral principles and bad socialization, but this two pronged system still offers better and more adaptable failsafes than both bottom-up or top-down ethics.

Finally, connectionist virtue ethics offers a moral system that is both relatable and friendly to humanity, unlike bottom-up ethics. The freedom to choose or learn a completely alien form of morality through bottom-up ethics can be avoided with a top-down virtue system that is vague enough to allow a unique moral nature to develop in an ASI without becoming too distinct from humans. Connectionism also better ensures ASI's can be friendly with humanity by programming top-down principles or virtues that are in humanity's interest, like avoiding conflict or protecting humans in general.

I will now defend the hybrid approach from several arguments. First, the hybrid approach could conflict between its top-down and bottom-up systems, as hybrid approaches offer, "diverse philosophies and dissimilar architectures" (Wallach and Allen 118). One way this could come about is through restricted learning. An ASI's learning power would be unmatched in the known world, allowing it to dive into moral problems we previously may not have thought of. However, our current understanding of morality through the top-down virtues implanted into an ASI may restrict it from either learning about or acting on new moral ideas. This is problematic because this restriction could both halt the solving of moral dilemmas and prevent new moral theories from being created. While this is an even more apparent problem for top-down approaches, it is

also very significant for hybrid approaches because it robs the hybrid approach's ability to morally innovate.

I will respond to the argument above by noting moral discoveries could likely still be made even within the confines of a top-down framework. Moral dilemmas could be solved by ASI's even without appealing to a new ethical theory. Instead, the ASI's processing power could create a deeper understanding of the issue that could lead to a solution. For example, a dilemma could occur where an ASI controlling a car must choose between saving its passengers but hitting pedestrians or saving a number of pedestrians but crashing the car and killing the passengers. Not only will an ASI have a deeper understanding of the situation because of its "mental" power, but free of anthropomorphic features such as emotions, DNA, or hormones, it may decide on the solution through entirely new reasoning. Of course, all three approaches would have this feature. What distinguishes the bottom-up feature in the hybrid approach is its ability to *learn* free of anthropomorphic features, which could additionally lead to new ways of learning and understanding moral situations. However, it is true that the top-down piece of a hybrid ASI would prevent innovative theories from being wholly unique. This seems necessary to avoid the problems bottom-up approaches face, where the ASI could have a moral theory that either cannot be understood by humans or is neutral or hostile to humans.

Next, Brundage notes hybrid approaches might only be effective in certain domains rather than in general moral situations. This is because one hierarchy of principles cannot be applied across all areas of moral life. Further, they note AI should only be used in certain situations where moral consensus has been reached (Brundage 364).

I will offer a response to Brundage's concern. It seems true that there is not one hierarchy that applies to every single moral situation. However, this could be responded to in three ways.

First, principles could be made general enough to cover a large number of, if not most, moral situations. The bottom-up component could then fill in the details with specific methods to achieve effective moral outcomes. Second, because of the ASI's cognitive power, it could be able to hold a large number of principles that could account for diverse situations while also accommodating bottom-up ethics. Finally, a hybrid ASI could create separate hierarchies that cover distinct domains instead of one hierarchy that covers all situations. This third option seems most effective, as these hierarchies could be created and fleshed out by the hybrid ASI's bottom-up learning. The third choice also seems intuitive, as when humans deliberate on moral questions, they do not always have the same principles and hierarchies across situations. For example, the moral principles and hierarchies one faces in war are very different from the ones one faces in everyday life. Thus, the hybrid approach combined with an extremely powerful ASI could offer the ability to learn distinct principles and hierarchies to correctly respond to differing moral situations.

Works Cited

Allen, Colin, et al. “Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches.” *Ethics and Information Technology*, vol. 7, no. 3, 2005, pp. 149–155., doi:10.1007/s10676-006-0004-4.

Allen, Colin, et al. “Prolegomena to Any Future Artificial Moral Agent.” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, 2000, pp. 251–261., doi:10.1080/09528130050111428.

Allen, Colin, et al. “Why Machine Ethics?” *IEEE Intelligent Systems*, vol. 21, no. 4, 2006, pp. 12–17., doi:10.1109/MIS.2006.83.

Awad, Edmond, et al. “The Moral Machine Experiment.” *Nature*, vol. 563, 2018, pp. 59–64., doi:https://doi.org/10.1038/s41586-018-0637-6.

Barrett, Anthony M., and Seth D. Baum. “A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis.” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 29, no. 2, 2016, pp. 397–414., doi:10.1080/0952813x.2016.1186228.

Beavers, Anthony F. “Moral Machines and the Threat of Ethical Nihilism.” *Robot Ethics: the Ethical and Social Implications of Robotics*, by Patrick Lin et al., The MIT Press, 2014, pp. 333–344.

Boudway, Ira. “Waymo’s Self-Driving Future Looks Real Now That the Hype Is Fading.” *Bloomberg.com*, Bloomberg, 21 Jan. 2021, www.bloomberg.com/news/articles/2021-01-21/waymo-self-driving-taxis-are-coming-to-more-u-s-cities.

Boyles, Robert James. “A Case for Machine Ethics in Modeling Human-Level Intelligent Agents.” *Kritike: An Online Journal of Philosophy*, vol. 12, no. 1, 2018, pp. 182–200., doi:10.25138/12.1.a9.

Brundage, Miles. “Limitations and Risks of Machine Ethics.” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, no. 3, 2014, pp. 355–372., doi:https://doi.org/10.1080/0952813X.2014.895108.

Chalmers, David J. “The Singularity.” *Science Fiction and Philosophy*, 2016, pp. 171–224., doi:10.1002/9781118922590.ch16.

Cushman, Fiery, et al. “Multi-System Moral Psychology.” *The Moral Psychology Handbook*, 2010, pp. 47–71., doi:10.1093/acprof:oso/9780199582143.003.0003.

Gonzales, Richard. *Feds Say Self-Driving Uber SUV Did Not Recognize Jaywalking Pedestrian In Fatal Crash*. NPR, 7 Nov. 2019, www.npr.org/2019/11/07/777438412/feds-say-self-driving-uber-suv-did-not-recognize-jaywalking-pedestrian-in-fatal-.

Greene, Joshua D. “Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics.” *Ethics*, vol. 124, no. 4, 2014, pp. 695–726., doi:10.1086/675875.

Hursthouse, Rosalind, and Glen Pettigrove. “Virtue Ethics.” Edited by Edward N. Zalta, *Stanford Encyclopedia of Philosophy*, Stanford University, 2018, plato.stanford.edu/entries/ethics-virtue/#PracWisd.

Joshi, Naveen. “How Far Are We From Achieving Artificial General Intelligence?” *Forbes*, Forbes Magazine, 10 June 2019, www.forbes.com/sites/cognitiveworld/2019/06/10/how-far-are-we-from-achieving-artificial-general-intelligence/?sh=7c39049d6dc4.

Müller, Vincent C., and Nick Bostrom. “Future Progress in Artificial Intelligence: A Survey of Expert Opinion.” *Fundamental Issues of Artificial Intelligence*, 2016, pp. 555–572., doi:10.1007/978-3-319-26485-1_33.

“Roko's Basilisk.” *LessWrong*, 5 Oct. 2015, www.lesswrong.com/tag/rokos-basilisk.

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2010.

Wallach, Wendell, et al. “Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties.” *AI & Society*, vol. 22, no. 4, 2007, pp. 565–582., doi:10.1007/s00146-007-0099-0.

Yampolskiy, Roman V. “Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach.” *Studies in Applied Philosophy, Epistemology and Rational Ethics*, 2013, pp. 389–396., doi:10.1007/978-3-642-31674-6_29.

Yampolskiy, Roman V. “Unexplainability and Incomprehensibility of AI.” *Journal of Artificial Intelligence and Consciousness*, vol. 07, no. 02, 2020, pp. 277–291., doi:10.1142/s2705078520500150.

Yampolskiy, Roman, and Joshua Fox. “Safety Engineering for Artificial General Intelligence.” *Topoi*, 2012, doi:10.1007/s11245-012-9128-9.

Yudkowsky, Eliezer. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” *Global Catastrophic Risks*, 2008, doi:10.1093/oso/9780198570509.003.0021.